主要内容

- 1.什么是数据预处理与表示
- **2.**描述性数据汇总
- **3. 结构数据**
  - **数据预处理（缺失值，噪音，重复数据，**数据变换，离散化，采样Sampling， 维归约及选择）
  - **数据表示（特征提取）**
- **4.**文本数据
  - **数据预处理**
  - **数据表示（特征提取）**
- **5.** 图像数据
  - **数据预处理**
  - **数据表示（特征提取）**

1. 定义

- 数据预处理是指在数据分析任务以前对数据进行的一些处理，以减小数据某些方面对算法性能的影响
  - 榨果汁-----切块
  - 分类----去噪音、缺失值填充、归一化等

- 特征选择是从原始特征数据集中选择出子集，是一种包含的关系，没有更改原始的特征空间。
- 特征提取是通过属性间的关系，如组合不同的属性得到新的属性，这样就改变了原来的特征空间。
- 特征表示是学习更高层的、具有语义等任务相关信息的表示

1. 数据预处理为什么是重要的？

- 没有高质量的数据，就没有高质量的挖掘结果
  - 高质量的决策必须依赖高质量的数据
    - e.g. 重复值或者空缺值将会产生不正确的或者令人误导的统计
  - 数据仓库需要对高质量的数据进行一致地集成
  - 低质量的数据 VS 算法性能
  - 低质数据挖掘是当前一个热门的研究方向

- 数据预处理将是构建数据仓库或者进行数据挖掘的工作中占工作量最大的一个步骤

1. 数据质量问题表现形式

- 具体问题表现为：
  - 误差：测量误差和收集误差或错误　　难以处理
  - 数据不一致

  - 噪声
    - 包含错误或者孤立点或离群点（outlier）
    - e.g. Salary = -10
    - 明显噪声和非明显噪声

“A, B, C”

  - 重复数据 duplication
    - 区分重复合法与否

  - 不完整
    - 缺少数据值；缺乏某些重要属性；仅包含汇总数据；
    - e.g., occupation=“”， weight= “ ” salary——学生NA

  - 其他问题：时效性，相关性， 采样合理性

2 描述性数据汇总

- 动机：为了更好的理解数据
  - 获得数据的总体印像
  - 识别数据的典型特征
  - 凸显噪声或离群点
  - 检查不一致数据，异常数据

- 度量数据的中心趋势
  - 均值、中位数、众数（模）、中列数

- 度量数据的离散程度
  - 四分位数、四分位数极差、方差等

度量的分类

- 度量可以分为三类：
  - 分布式度量(distributive measure)：将函数用于n个聚集值得到的结果和将函数用于所有数据得到的结果一样
    - 比如：count()，sum()，min()，max()等
  - 代数度量(algebraic)：可以 通过在一个或多个分布式度量上应用一个代数函数而得到
    - 比如：平均值函数avg()（ avg() =sum()/count()）
  - 整体度量(holistic)：必须对整个数据集计算的度量
    - 比如：median()，mode()，rank()

度量中心趋势 (1)

- 算术平均值 $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 加权算术平均

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- 截断均值（trimmed mean）：去掉高、低极端值得到的均值
  - e.g. 计算平均工资时，可以截掉上下各2%的值后计算均值，以抵消少数极端值的影响

度量中心趋势 (1)

- 中位数：有序集的中间值或者中间两个值平均
  - 整体度量；但是可以通过插值法计算近似值

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}}\right) width$$

L_1 ：中位数区间的下界

N ：整个数据集中值的个数

$(\sum freq)_l$ ：低于中位数区间的所有区间的频率和

$freq_{median}$ ：中位数区间的频率

$width$ ：中位数区间的宽度

度量中心趋势 (2)

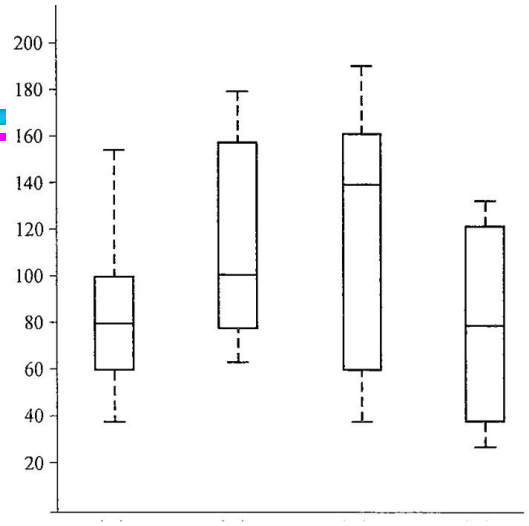
- 众数（Mode，也叫*模*）：集合中出现频率最高的值
  - 单峰的（unimodal，也叫单模态）、双峰的（bimodal）、三峰的（trimodal）；多峰的（multimodal）
  - 对于适度倾斜（非对称的）的单峰频率曲线，可以使用以下经验公式计算众数

$$mean - mode = 3 \times (mean - median)$$

度量数据的离散度 （1）

- 最常用度量：极差、五数概括（基于四分位数）、中间四分位数极差和标准差
  - 极差（range）：数据集的最大值和最小值之差
  - 百分位数*(percentile)*：第*k*个百分位数是具有如下性质的值*x*：*k%*的数据项位于或低于*x*
    - 中位数就是第50个百分位数
  - 四分位数：Q_1 *(25th percentile)*, Q_3 *(75th percentile)*
  - 中间四分位数极差*(IQR)*： $IQR = Q_3 - Q_1$
  - 孤立点：通常我们认为：挑出落在至少高于第三个四分位数或低于第一个四分位数 *1.5*×*IQR*处的值

度量数据的离散度 （2）



- 五数概括: *min*, Q_1, *Median*, Q_3, *max*
- 盒图：数据分布的一种直观表示
- 方差和标准差
  - 方差s^2：*n*个观测之$x_1, x_2 \ldots x_n$的方差是

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

  - 标准差*s*是方差s^2的平方根
    - 标准差s是关于平均值的离散的度量，因此仅当选平均值做中心度量时使用
    - 所有观测值相同则 s＝0，否则 s>0
    - 方差和标准差都是代数度量

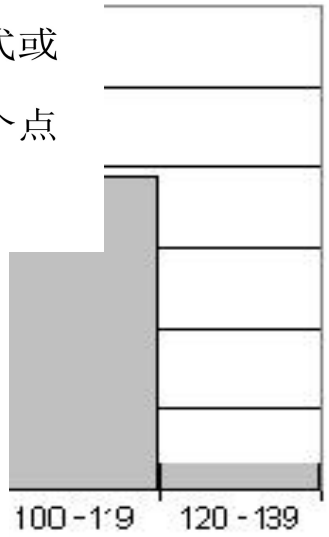
基本统计类描述的图形显示——直方图

- 常用的显示数据汇总和分布的方法：
  - 直方图、分位数图、q-q图、散布图和局部回归曲线
- 直方图：一种单变量图形表示方法
  - 将数据分布划分成不相交的子集或桶，通常每个桶宽度一致并用一个矩形表示，其高度表示桶中数据在给定数据中出现的计数或频率

loess

散布图

- 确定两个量化的变量之间看上去是否有联系、模式或者趋势的最有效的图形方法之一
- 散布图中的每个值都被视作代数坐标对，作为一个点画在平面上
- 易于观察双变量数据在平面上的分布



loess曲线

- loess曲线为散布图添加一条平滑的曲线，以便更好的观察两个变量间的依赖模式
- Loess (local regression)意指“局部回归”，为了拟合loess曲线，需要两个参数：平滑参数α，被回归拟合的多项式的阶λ

2 结构化数据预处理

- 数据清理
  - 填写空缺的值，平滑噪声数据，识别、删除孤立点，解决不一致性
- 数据集成
  - 集成多个数据库、数据立方体或文件
- 数据变换
  - 规范化和聚集
- 数据归约
  - 得到数据集的压缩表示，它小得多，但可以得到相同或相近的结果
- 数据离散化
  - 数据归约的一部分，通过概念分层和数据的离散化来规约数据，对数字型数据特别重要

2 结构化数据预处理——空缺值处理

- 数据并不总是完整的
  - 数据库表中，很多条记录的对应字段没有相应值，比如销售表中的顾客收入
- 引起空缺值的原因
  - 设备异常
  - 与其他已有数据不一致而被删除
  - 因为误解而没有被输入的数据
  - 在输入时，有些数据因为得不到重视而没有被输入
  - 对数据的改变没有进行日志记载
- 空缺值要经过推断而补上

2 结构化数据预处理——空缺值处理

- 忽略元组：当类标号缺少时通常这么做（假定挖掘任务设计分类或描述），当每个属性缺少值的百分比变化很大时，它的效果非常差。
- 人工填写空缺值：工作量大，可行性低

- 填充：基于对已知的取值的观察和分析，对缺失值进行自动填充
  - 用一个全局变量填充空缺值：比如使用unknown或∞
  - 使用属性的（加权/截断）平均值/众数/中位数填充空缺值
  - **基于Bayesian，knn或判定树推断的方法**

主观题 10分

设置

1）分析数据分布和关系，利用众数或均值填充其中的缺失值，并给出填充计算方法或策略
2）试分析这种填充方法的特点

Age	sex	region	income	married	children	car	label
40,	MALE,	TOWN,	30085.1,	YES,	3,	YES,	NO
51,	FEMALE,	INNER,	16575.4,	YES,	0,	YES,	NO
23,	FEMALE,	TOWN,	20375.4,	?,	3,	NO,	NO
57,	FEMALE,	RURAL,	50576.3,	YES,	0,	NO,	NO
57,	FEMALE,	TOWN,	37869.6,	YES,	2,	NO,	YES
22,	MALE,	RURAL,	8877.07,	NO,	0,	NO,	YES
58,	MALE,	TOWN,	24946.6,	YES,	0,	YES,	NO

正常使用主观题需2.0以上版本雨课堂

作答

2 结构化数据预处理——空缺值处理

- **插值方法有Hermite插值、分段插值、样条插值法，而最主要的有拉格朗日插值法和牛顿插值法。**
- **拉格朗日插值法**

  - 噪声：一个测量变量中的随机错误或偏差
  - 引起不正确属性值的原因
    - 数据收集工具的问题
    - 数据输入错误
    - 数据传输错误
    - 技术限制
    - 命名规则的不一致
  - 其它需要数据清理的数据问题
    - 重复记录
    - 不完整的数据
    - 不一致的数据

$L(x)$

2 结构化数据预处理——噪声数据

- 分箱(binning):
  - 首先排序数据，并将他们分到等深的箱中
  - 然后可以按箱的平均值平滑、按箱中值平滑、按箱的边界平滑等等
- 回归
  - 通过让数据适应回归函数来平滑数据
- 聚类:
  - 监测并且去除孤立点
- 计算机和人工检查结合
  - 计算机检测可疑数据，然后对它们进行人工判断

2 结构化数据预处理——规范化

- 最小－最大规范化

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score规范化
  - 最大最小值未知，或者离群点影响较大的时候适用

$$v' = \frac{v - mean_A}{standard\_dev_A}$$

- 小数定标规范化 $v' = \frac{v}{10^j}$

其中，j是使 Max(| v'|)<1的最小整数

主观题 10分

设置

分析age和income数值分布，并选择合适的规范化方法进行规范化

Age	sex	region	income	married	children	car	label
40,	MALE,	TOWN,	30085.1,	YES,	3,	YES,	NO
51,	FEMALE,	INNER,	16575.4,	YES,	0,	YES,	NO
23,	FEMALE,	TOWN,	20375.4,	?,	3,	NO,	NO
57,	FEMALE,	RURAL,	50576.3,	YES,	0,	NO,	NO
57,	FEMALE,	TOWN,	37869.6,	YES,	2,	NO,	YES
22,	MALE,	RURAL,	8877.07,	NO,	0,	NO,	YES
58,	MALE,	TOWN,	24946.6,	YES,	0,	YES,	NO

答案示例：

Age={0.1，0.2，0.3……0.3}

income={0.1，0.2，0.3……0.3}

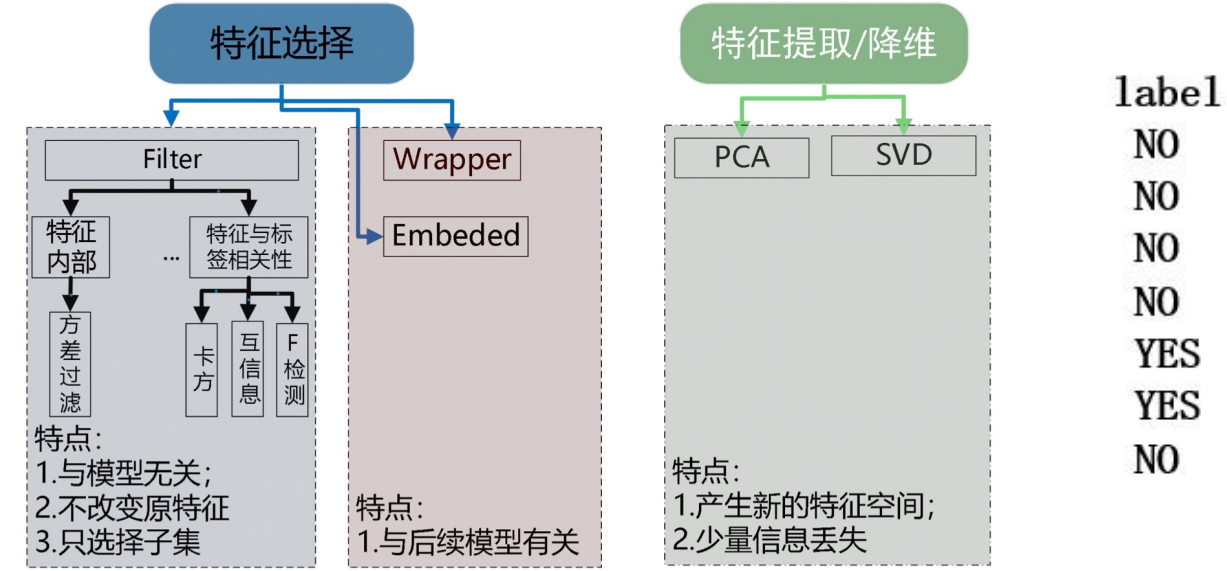
正常使用主观题需2.0以上版本雨课堂

作答

2 结构化数据预处理——特征选择、提取和表示

- 特征的处理分为三大类：选择、提取和表示。其目的是降维和提升对任务的表达能力

- 属性降维
  - 特征选择
  - 特征提取



label
NO
NO
NO
NO
YES
YES
NO

- 属性构造：通过现有属性构造新的属性，并添加到属性集中；以增加对高维数据的结构的理解和精确度

2 结构化数据预处理——特征选择、提取和表示

- 特征选择通过删除不相干的属性或维减少数据维度，一般是基于原始特征空间
  - 冗余特征 income--tax
  - 不相关特征　id
- 特征选择的方案
  - embedded
  - filter
  - wrapper
- 特殊的特征选择——特征加权
  - 找出最小属性集，使得数据类的概率分布尽可能的接近使用所有属性的原分布
  - 减少出现在发现模式上的属性的数目，使得模式更易于理解

维度增加不仅仅是数据体量上的增加，往往伴随着稀疏性的增加

2 结构化数据预处理——特征选择、提取和表示

- 特征选择概述
  - **按照最优特征集合的产生过程**
    - **穷举法**：是遍历所有特征子集选取最优集的方法，虽然这样做一定可以找到最优解；
      - ✓ 例如：回溯法等
    - **启发式方法**：利用某种主观的启发式规则进行搜索的过程，从而避免搜索全部子集产生的庞大时空开销；
      - ✓ 例如：著名的C4.5决策树方法就是采用信息增益作为启发式规则来获取近似最优解；
    - **随机式方法**：完全随机法和概率随机法
      - ✓ 例如：遗传算法就是通过将某个子集放在具体环境里不断交叉、复制等产生更适应环境的解；

2 结构化数据预处理——特征选择、提取和表示

- 特征选择概述
  - 按照基于分类问题的特征降维技术分类
    - **过滤器**（Filters）：根据统计指标值对所有特征进行排序，取得前*k*个值；或者设置选择阈值*h*；
      - **方差**
      - **OR(Odds Ratio)**
      - **信息增益等指标**
      - **...**
    - 计算简单，与模型无关
    - **忽略了特征空间子集在解决分类问题中的作用**

2 结构化数据预处理——特征选择、提取和表示

- 过滤特征选择的指标
- 1)相似性，相异性
  - 欧几里得距离
  - cos距离

$$\cos(x, y) = \frac{x.y}{\|x\|\|y\|}$$

例1：两个文档相似度

X=（3，2，0，5，0，0，0，2，0，0）

Y=（1，0，0，0，0，0，0，1，0，2）

$$x.y = \sum x_i * y_i = 5$$

$$\|x\| = \sqrt{\sum x_i * x_i} = 6.48$$

$$\|y\| = 2.45$$

2 结构化数据预处理——特征选择、提取和表示

- 相似性，相异性

X=（1，0，0，0，0，0，0，0）
Y=（1，0，1，0，0，0，0，1）

  - **简单匹配系数**

$$SMC = \frac{\text{值匹配的属性个数}}{\text{属性个数}} = \frac{f_{11} + f_{00}}{f_{11} + f_{01} + f_{10} + f_{00}}$$

  - **Jaccard系数**

$$J = \frac{f_{11}}{f_{11} + f_{01} + f_{10}}$$

  - **广义Jaccard（Tanimoto）**

$$EJ(x, y) = \frac{x.y}{\|x\|^2 + \|y\|^2 - x.y}$$

2 结构化数据预处理——特征选择、提取和表示

- 相关性
  - **皮尔森系数Pearson's correla**

$$corr(x, y) = \frac{s_{xy}}{s_x s_y}$$

- 协方差:

$$s_{xy} = \text{covariance}(x, y) = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})$$

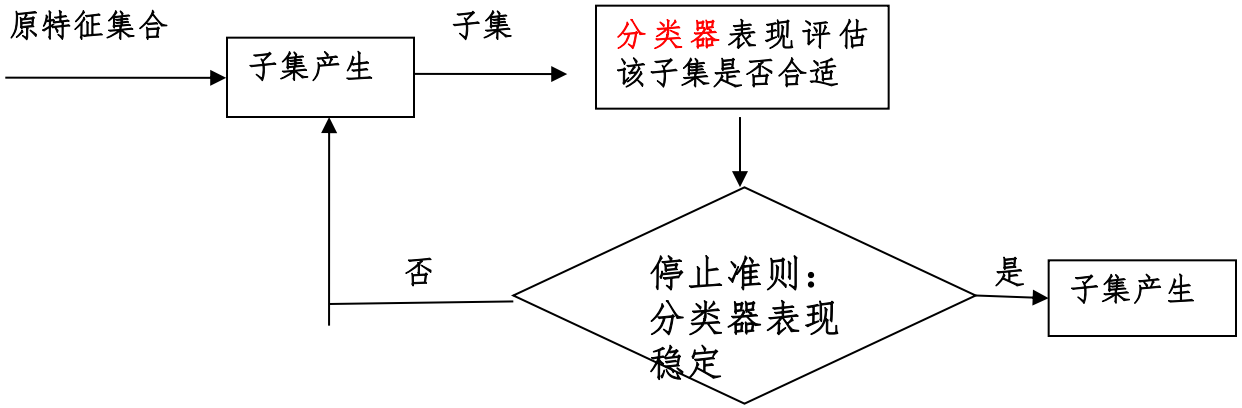
- 标准差:

$$s_x = \text{standard\_deviation}(x) = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

- 相关性
  - 正相关 corr=1　　x=(1,-2) y=(2,-4)
  - 负相关 corr=-1　　x=(1,-2) y=(-2,4)
  - 非线性相关 corr=0　x=(1,-2) y=(-2,-4)

2 结构化数据预处理——特征选择、提取和表示

- 特征选择--
  - **包装器**（Wrappers）：选择一个基分类器，启发式搜索所有可能的特征子空间，评估分类效果，选择/过滤具有代表性的特征组合空间；



  - **由于要计算每个特征的分类性能，计算量较大；**
  - 所使用的评估算法未知，可能对评估算法敏感
  - 子集产生不依赖算法
  - sklearn.feature_selection中的RFE

2 结构化数据预处理——特征选择、提取和表示

- 特征选择--
  - **嵌入**（Embeded）：类似于Filter，算法决定产生子集。先使用算法和模型进行训练，得到各个特征的权值系数，根据权值系数从大到小选择特征。

- 特征提取，也叫属性抽取——将高维特征空间映射到低维空间
  - 高维空间中的特征是原特征空间中没有的新的特征
  - 这些新特征大多为组合变换后的特征

    - 典型方法
      - PCA（Principal Components Analysis）
      - SVD（Singular Value Decomposition）

2 结构化数据预处理——特征选择、提取和表示

- PCA（Principal Components Analysis）——非监督

  一种用于连续属性的线性代数技术，它找出新的属性（主成分），这些属性是原属性的线性组合，是相互正交的，并且捕获数据的最大变差，最小协方差（协方差为0，相互独立）。PCA是一种主要的特征降维技术。

PCA降维准则：

（1） 最近重构性：重构后的点距离原来的点的误差之和最小。

（2） 最大可分性：样本在低维空间的投影尽可能分开

（3）不相关：将可能存在相关性的变量数据转换为一组线性不相关的变量

PCA本质上是将方差最大的方向作为主要特征，并且在各个正交方向上将数据“零相关”，也就是让它们在不同正交方向上没有相关性。

输入：训练样本集 $D=x^{(1)},x^{(2)},\ldots,x^{(m)}$,低维空间维数 d' ;

过程：.

1: 对所有样本进行中心化（去均值操作）： $x_j^{(i)} \leftarrow x_j^{(i)} - \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}$;

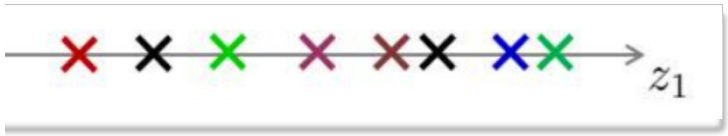
2: 计算样本的协方差矩阵 XX^T ;

3: 对协方差矩阵 XX^T 做特征值分解 ;

4: 取最大的 d' 个特征值所对应的特征向量 $w_1,w_2,\ldots,w_{d'}$

5: 将原样本矩阵与投影矩阵相乘： $X\cdot W$ 即为降维后数据集 X' 。其中 X 为 $m\times n$ 维，$W=[w_1,w_2,\ldots,w_{d'}]$ 为 $n\times d'$ 维。

5: 输出：降维后的数据集 X'



2 结构化数据预处理——特征选择、提取和表示

1、去中心化、标准化
2、计算协方差矩阵
3、特征值分解
4、降维后的数据

去中心化的好处：
1、方便协方差求解
2、特征值分解

- PCA算法优点：
  - 使得数据集更易使用；降低算法的计算开销；
  - 去除噪声；　使得结果容易理解；
  - 完全无参数限制、无监督。
- PCA算法缺点：
  - 难以利用先验知识来干预方法，可能会得不到预期的效果，效率也不高；
  - 特征值分解有一些局限性，如变换的矩阵必须是方阵；
  - 在非高斯分布情况下，PCA方法得出的主元可能不是最优

元)	面积(百平米)
	4.4
	−1.6
	−2.6
	1.9
	−2.1

去中心化的好处：
1、方便协方差求解
2、特征值分解

2 结构化数据预处理——特征选择、提取和表示

1、去中心化、标准化
2、计算协方差矩阵
3、特征值分解
4、降维后的数据

	房价(百万元)	面积(百平米)
a	10	9
b	2	3
c	1	2
d	7	6.5
e	3	2.5

中心化 →

	房价(百万元)	面积(百平米)
a	5.4	4.4
b	−2.6	−1.6
c	−3.6	−2.6
d	2.4	1.9
e	−1.6	−2.1

去中心化的好处：
1、方便协方差求解
2、特征值分解